Ranking and selecting association rules based on dominance relationship

Slim Bouker

Ranking and selecting association rules based on dominance relationship

Slim Bouker and **Rabie Saidi** and **Sadok Ben Yahia** and **Engelbert Mephu Nguifo** [1]

Abstract. The huge number of association rules represent the main obstacle that a decision maker faces. In order to bypass this obstacle, an efficient selection of rules must be performed. Since selection is necessarily based on evaluation, many interestingness measures have been proposed. However, the abundance of these measures caused a new problem which is the heterogeneity of the evaluation results and this created confusion to the decision. In this scope, we propose a novel approach to discover interesting association rules without favouring or excluding any measure by adopting the notion of dominance between rules. Our approach bypasses the problem of measure heterogeneity and find a compromise between their evaluations and also bypasses another non-trivial problem which is the threshold value specification.

1 INTRODUCTION

Mining association rules is one of the core tasks in data mining research. Since its first formalization in [1], the research on association rules has become very popular among the data mining researchers, as it provides an opportunity to extract relevant and valuable relationship between attributes in transaction databases.

At present, association rules are widely used in the *decision making* related to various areas such as telecommunication networks, market and risk management, inventory control etc, where the databases are generally large [13]. However, it is well known that data mining algorithms produce huge numbers of rules [8]. Hence, the decision maker is unable to determine the most interesting ones and consequently unable to make decisions. In order to face this obstacle, an efficient evaluation of rules has become a need rather than being a rational choice. Several works have been devoted to study the interestingness of association rules [6], [7], [17], [19]. As a consequence, a panoply of statistical measures, obeying different semantics, have been proposed. Although these measures allow evaluating rules from various sights, yet their abundance (≈ 60) has yielded another problem for the decision maker. Indeed, the outputs of evaluations vary from a measure to another and may even be contradictory since the measures evaluate differently the rules under consideration. That is why, it is common that a given rule be considered relevant according to a measure and irrelevant according to another.

The problem caused by the abundance of measures has led to a trend of works that focuss on proposing approaches to assist the user(*i.e.,* the decision maker) in selecting the measures qualified to be the most adequate to the decision scope. These approaches can be classified into two main categories namely the expert-based approaches and the property-based approaches. In the first category, different studies have compared the ranking of rules by human experts to the ranking of rules by various measures. Then, they suggested choosing the measure that produces the ranking which most resembles the expert one [15], [18]. These studies were based on specific datasets and experts. Thus, their results cannot be taken as general conclusions. Moreover, in a real problem, it is not always possible to get rule ranking by experts. As for the second category, to reduce the number of measures, many properties have been reported in [4]. Geng and Hamilton surveyed the interestingness of measures and summarized nine properties to address that issue. Using properties facilitates a general and practical way to automatically identify interesting measures. This trend has been enriched by different other works [2], [5], [11], [12] with an extensive number of properties. Nevertheless, these properties are not standards [10]. Hence, they do not guarantee selecting only one best measure. Indeed, a wide range of UCI[2] datasets were also used to study the impact of different properties. The results show no single measure can be introduced as an obvious winner [5]. Then, in the case of selecting many measures, the problem related to the variety of outputs, mentioned above, persists. In other words, the user cannot proceed towards a unique selection of rules. Whatever one measure is selected or more, nothing guarantees that they are the "best" ones and some better suited measures may be excluded for the simple reason that the used properties do not take into account the specificity of decision context.

Our contribution lies within this scope. In this paper, we propose a novel approach to discover interesting association rules without favoring or excluding any mesure among the used measures. For this purpose, we integrate into the rule selection process, the *skyline operator* [3] whose fundamental principle relies on the notion of *dominance*. Skyline operator is used to resolve mathematical and economics problems such as maximum vectors [9], Pareto set [14] and multi-objective optimization [16]. In our work, the skyline operator comprises the rules that are supposed to be the most interesting ones while taking into account several measures. The dominance relationship which is the corner stone of the skyline operator is applied on rules and can be presented as follows: a rule r is said dominated by another rule r', if for all used measures, r is less relevant than r'. The former rule (*i.e., r*) is discarded from the result, not because it is not relevant for one of the mesure but because it is not interesting according to the combination of all measures. Our approach bypasses the problem of measure selection by finding a compromise between the different outputs and also bypasses another nontrivial problem which is the threshold value specification.

The remainder of this paper is organized as follows. Section 2 gives a brief definitions related to association rules and introduce

[1] Clermont Universit, Universit Blaise Pascal, LIMOS, BP 10448, F-63000 CLERMONT-FERRAND, email: slim.bouker@isima.fr, saidi@isima.fr, sadok.benyahia@fst.rnu.tn, mephu@isima.fr

[2] http://archive.ics.uci.edu/ml/

the dominance relationship. We propose and detail our approach of rule selection in section 3. An extension of our approach to enable rule ranking is presented in section 4. Concluding points and some prospectives make the body of section 5.

2 ASSOCIATION RULES AND DOMINANCE RELATIONSHIP

In this section we first recall basic definitions related to association rules. Then, we present these rules as numeric vectors within the same dimension after having been evaluated by a set of measures. This vector format, allows us to benefit from the concept of *dominance* and adapt it to our scope as described in section 2.2.

2.1 Association rules

Let $\mathcal{I}$ be a set of literal called items, an itemset corresponds to a non null subset of $\mathcal{I}$. These itemsets are gathered together in the set $\mathcal{L}$: $\mathcal{L} = 2^{\mathcal{I}} \setminus \emptyset$. In a transactional dataset, each transaction contains an itemset of $\mathcal{L}$. Table 1(a) presents a transactional dataset $\mathcal{D}$ where 10 transactions denoted by $t_1, \ldots, t_{10}$ are described by 4 items denoted by a, b, c, d. The support of an itemset X, denoted $supp$(X), is the number of transactions containing X. The negative support $supp(\overline{X})$ is the number of transactions that do not contain X.

An association rule r is a relation between itemsets of the form r: X→Y where X and Y are itemsets, and $X \cap Y = \emptyset$. Itemsets X and Y are called, respectively, premise and conclusion of r. The support of r is equal to the number of transactions containing both X and Y, $supp(r) = supp(X \cup Y)$. We notice that interesting measures for association rules are usually defined using support counts as presented in Table 1(b).

	a	b	c	d
t_1			×	×
t_2	×			
t_3	×			×
t_4			×	
t_5		×		×
t_6	×			×
t_7			×	
t_8				×
t_9		×	×	
t_{10}			×	×

(a) *A transaction data set* $\mathcal{D}$

Rule	*Freq*	*Conf*	*Pearl*
r_1: $a \rightarrow d$	0.20	0.67	0.02
r_2: $b \rightarrow c$	0.10	0.50	0.00
r_3: $b \rightarrow d$	0.10	0.50	0.02
r_4: $c \rightarrow d$	0.20	0.40	0.10
r_5: $d \rightarrow a$	0.20	0.33	0.02
r_6: $d \rightarrow c$	0.20	0.33	0.10
r_7: $c \rightarrow b$	0.10	0.20	0.01
r_8: $d \rightarrow b$	0.10	0.17	0.02

(b) *A table rlation* $\Omega(\mathcal{R}, \mathcal{M})$

Name	*Definition*	*Domain*
Frequency	$\frac{supp(X \cup Y)}{\lvert D \rvert}$	[0, 1]
Confidance	$\frac{supp(X \cup Y)}{supp(X)}$	[0, 1]
Pearl	$\frac{supp(X)}{\lvert D \rvert} \times \lvert \frac{supp(X \cup Y)}{supp(X)} - \frac{supp(Y)}{\lvert D \rvert} \rvert$	[0, 1]

(c) *Some measures of* $\mathcal{M}$

Table 1. Example of a dataset transaction and measures.

2.2 Dominance relationship

After mining association rules from transactional dataset $\mathcal{D}$ (*e.g.,* Table1(a)), a set $\mathcal{R}$ of rules is obtained (*e.g.,* Table1(b) first column). Rules of $\mathcal{R}$ are evaluated by a set $\mathcal{M}$ of measures (*e.g.,* Table1(c)) to form a relational table Ω (*e.g.,* Table1(b)). Formally, $\Omega = (\mathcal{R}, \mathcal{M})$ with the set $\mathcal{M} = \{m_1, \ldots, m_k\}$ of measures as attributes and the set $\mathcal{R} = \{r_1, \ldots, r_n\}$ of rules as objects. We note by $r[m]$ the value of the measure m for the rule r, $r \in \mathcal{R}$ and $m \in \mathcal{M}$. Since the evaluation of rules varies from a measure to another, using several measures could lead to different outputs (relevant rules with respect to a measure). For example, r_1, and r_2 are the best two rules according to the evaluation of the *Confidence* measure whereas it is not the case according to the evaluation of *Pearl* measure which favors r_4 and r_6. This difference of evaluation is confusing for any process of rule selection or ranking. Other examples can be found in Table1(b).

Based on the above formulation of Ω, we can utilize the notion of dominance between rules to address their ranking as well as the selection of relevant ones. Before, formulating the dominance relationship between rules we need to define it at the level of measure values. To do that, we define value dominance as follows:

Definition 1 ***(Value Dominance)*** *Given two values of a measure m corresponding to two rules r and r', we say that $r[m]$ dominates $r'[m]$, noted by $r[m] \succeq r'[m]$, iff $r[m]$ is preferred to $r'[m]$. If $r[m] \succeq r'[m]$ and $r[m] \neq r'[m]$ then we say that $r[m]$ strictly dominates $r'[m]$, we note $r[m] \succ r'[m]$.*

Remark. The preference between two values differs from a measure to another.
Example. Given v and v' two values and m, $m' \in \mathcal{M}$ two measures, such that the best values in the domain of m and the domain of m' are respectively 0 and 1. For instance, if v = 0.3 and v' = 0.8, then v strictly dominates v' with respect to m, whereas v' strictly dominates v with respect to m'.

To make the dominance relationship scale to the level of rules, we give the following definition:

Definition 2 ***(Rule Dominance)*** *Given two rules r, $r' \in \mathcal{R}$, the dominance relationship according to the set of measures $\mathcal{M}$ is defined as follows:*

- *r dominates r', noted $r \succeq r'$, iff $r[m] \succeq r'[m]$, $\forall\, m \in \mathcal{M}$.*
- *If $r \succeq r'$ and $r' \succeq r$, i.e., $r[m] = r'[m]$, $\forall\, m \in \mathcal{M}$ then r and r' are said equivalent, we note $r \equiv r'$.*
- *If $r \succeq r'$ and $\exists\, m \in \mathcal{M}$ such that $r[m] \succ r[m]$, then r' is strictly dominated by r and we note $r \succ r'$.*

It is easy to verify that strict dominance relationship is:

- irreflexive: $r \not\succ r$, *i.e,* $r \succ r$ is false for each $m \in \mathcal{M}$,
- transitive: $\forall\, r$, r' and $r'' \in \mathcal{R}$, if $r \succeq r'$ and $r' \succeq r''$ then $r \succeq r''$.

Example. Given the relation table Ω in Table1(b), the rule r_3 strictly dominates r_2 because $r_3[Freq] \succeq r_2[Freq]$, $r_3[Conf] \succeq r_2[Conf]$ and $r_3[Pearl] \succ r_2[Pearl]$.

When a rule r dominates another rule r' with respect to $\mathcal{M}$, this means that r is equivalent or better than r' for all measures. Indeed, the values of r dominate those of r' for all measures. The dominance relationship allows comparing two rules with respect to all measures at the same time. Hence, it can be used to bypass the problem of difference of evaluations. The rules which are dominated by others (at least one) according to $\mathcal{M}$ are not relevant and must be eliminated. The skyline operator for association rules formalizes this intuition.

Definition 3 ***(Skyline operator)*** *The skyline of Ω over $\mathcal{M}$, denoted by $Sky_M(\Omega)$, is the set of rules from Ω defined as follows:*

$$Sky_M(\Omega) = \{\, r \in \mathcal{R} \mid \nexists\, r' \in \mathcal{R},\, r' \succ r \}$$

In other words, the skyline of Ω is the set of undominated rules of $\mathcal{R}$ according to $\mathcal{M}$. For instance, from the relation table Ω in Table1(b), $Sky_M(\Omega) = \{r_1, r_4\}$ because there is no rule in $\mathcal{R}$ which dominates r_1 or r_4.

3 DISCOVERING UNDOMINATED RULES

In this section, we describe our approach to discover the undominated rules. In the next subsection, we introduce the necessary formalization that helps with the generation of the undominated rules. Based on this formalization, we propose SKYRULE, the algorithm meant to concretize the skyline operator.

3.1 Formalization

To discover the undominated rules, a naïve approach consists in comparing each rule with all other ones. However, association rules are often present in huge number which make it very costly to perform pairwise comparisons. In the following, we show how to remedy this problem. First, we define the reference rule.

Definition 4 ***(Reference Rule)*** *A reference rule $r^{\perp}$ is a fictitious rule that dominates all the rules of $\mathcal{R}$. Formally: $\forall\, r \in \mathcal{R}$, $r^{\perp} \succeq r$.*

Example. From the relational table Ω in Table1, we can consider $r^{\perp}$ as the fictitious rule such that for each measure $m \in \mathcal{M}$, $r^{\perp}[m]$ is the maximum value appearing in the active domain of m, then $r^{\perp} = \langle 0.2, 0.67, 0.10 \rangle$. Hence, there is no rule in $\mathcal{R}$ that dominates $r^{\perp}$.

In practice, measures are heterogenous and defined within different domains. For our purpose, $\mathcal{M}$ must be normalized into $\widehat{\mathcal{M}}$ within one interval [p,q]. In other words, each measure $m \in \mathcal{M}$ must be normalized into $\widehat{m} \in \widehat{\mathcal{M}}$ within [p,q]. The normalization of a given measure m is performed depending on its domain and the statistical distribution of its active domain. We recall that the active domain of a measure m is the set of its values in Ω. The normalization is a statistical problem that we are not dealing with. Obviously, The normalization of a measure do not modify the domination relationship between two given values.

Definition 5 ***(Degree of similarity)*** *Given two rules $r,\, r' \in \mathcal{R}$, the degree of similarity between r and r' with respect to $\widehat{\mathcal{M}}$ is defined as follows:*

$$DegSim(r, r') = \frac{\sum_{i=1}^{k} \mid r[\widehat{m}_i] - r'[\widehat{m}_i] \mid}{k}$$

with $\mid x - y \mid$ is the absolute value of $(x - y)$, x and y $\in$ [p,q].

Example. Let's consider our running example using the relation table Ω in Table1(b). Since all measures are defined within the same domain [0,1], we can compute, without normalization, the degree of similarity between each rule and the reference rule given in the previous example. $DegSim(r^{\perp},r_1)$ = 0.02, $DegSim(r^{\perp},r_2)$ = 0.12, $DegSim(r^{\perp},r_3)$ = 0.11, $DegSim(r^{\perp},r_4)$ = 0.09, $DegSim(r^{\perp},r_5)$ = 0.14, $DegSim(r^{\perp},r_6)$ = 0.11, $DegSim(r^{\perp},r_7)$ = 0.22, $DegSim(r^{\perp},r_8)$ = 0.23.

After giving the necessary definitions (reference rule and degree of similarity), the following lemma gives a remedy to the issue evoked in the beginning of section 3.1. Indeed, it offers a rapid solution rather than pairwise comparisons; to find undominated rules.

Lemma 1 *Let $r \in \mathcal{R}$ be a rule having the minimal degree of similarity with $r^{\perp}$, then $r \in Sky_M(\Omega)$.*

Proof 1 *Let $r \in \mathcal{R}$ be a rule having the minimal degree of similarity with $r^{\perp}$ and we suppose that $r \notin Sky_M(\Omega)$, then there exist a rule $r' \in \mathcal{R}$ that strictly dominates r, which means that $\forall\, m \in \mathcal{M}$, $r'[m] \succeq r[m]$ and $\exists\, m' \in \mathcal{M}$, $r'[m'] \succ r[m']$. Hence, $DegSim(r^{\perp},r') < DegSim(r^{\perp},r)$ which is absurd since r has the minimal degree of similarity with $r^{\perp}$.*

After identifying an undominant rule r, the rules dominated by r must be identified by comparing them to r. Naïvely, r must be compared to all rules in $\mathcal{R}$, yet we show in the following that we can reduce the set of rules to be compared with r into a subset of $\mathcal{R}$.

Definition 6 ***(undominated space)*** *Let r be an undominated rule. If there exists a rule r' which is not dominated by r such that $r \not\equiv r'$, then there exists at least a measure $m \in \mathcal{M}$ such that $r'[m] \succ r[m]$. Since there exist k measures in $\mathcal{M}$, then there are k sets such that each one of them may contain rules not dominated by r. For each measure $m_i \in \mathcal{M}$, i=1...k, the corresponding set s_i^r of rules not dominated by r is defined as follows:*

$$s_i^r = \{\, r' \in \mathcal{R} \mid r \not\succ r' \text{ and } r'[m_i] \succ r\,[m_i] \}$$

These k sets compose the undominated space of r, noted $\mathcal{S}^r = \{s_i^r\}$, i=1...k.

Example. From our toy example presented in Table1, for the undominated rule r_1, $s_1^{r_1} = \emptyset$, $s_2^{r_1} = \emptyset$ and $s_3^{r_1} = \{r_4, r_6\}$. $s_1^{r_1}$ and $s_2^{r_1}$ are empty because there is no rule $r \in \mathcal{R}$ such that $r[m_1] \succ r_1[m_1]$ or $r[m_2] \succ r_1[m_2]$. However, $s_3^{r_1}$ contain r_4 and r_6 because $r_4[m_3] \succ r_1[m_3]$ and $r_6[m_3] \succ r_1[m_3]$. Following a similar reasoning, for the undominated rule r_4, $s_1^{r_4} = \emptyset$, $s_2^{r_4} = \{r_1, r_2, r_3\}$ and $s_3^{r_4} = \emptyset$.

Lemma 2 *Let $r, r' \in \mathcal{R}$ be two undominated rules and $s^r \in \mathcal{S}^r$. If $r' \notin s^r$ then $\forall\, r'' \in s^r$, $r' \not\succ r''$.*

Proof 2 *Given $r,\, r' \in \mathcal{R}$ two undominated rules and $s^r \in \mathcal{S}^r$ corresponding to a measure $m \in \mathcal{M}$. If $r' \notin s^r$, then $r'[m] \not\succ r[m]$ which means $r[m] \succeq r'[m]$ (1). Moreover, since $r'' \in s^r$ then $r''[m] \succ r[m]$ (2). According to the dominance transitivity, (1) and (2) mean $r''[m] \succ r'[m]$. Hence, $r' \not\succ r''$.*

Lemma 3 *Let be $r,\, r' \in \mathcal{R}$ and $s^r \in \mathcal{S}^r$ such that r is an undominated rule and $r' \in s^r$. If r' has the minimal degree of similarity with $r^{\perp}$ among the rules in s^r, then $r' \in Sky_M(\Omega)$.*

Proof 3 *Given $r,\, r' \in \mathcal{R}$ and $s^r \in \mathcal{S}^r$ such that $r' \in s^r$ and r' has the minimal degree of similarity with $r^{\perp}$ among the rules in s^r. Suppose that $r' \notin Sky_M(\Omega)$, that means there exists a rule $r'' \in \mathcal{R}$ such that $r'' \succ r'$. According to lemma 2, r'' must be in s^r since any rule not belonging to s^r cannot dominate r'. Moreover $\forall\, m \in \mathcal{M}$, $r''[m] \succeq r'[m]$ and $\exists\, m' \in \mathcal{M}$, $r''[m'] \succ r'[m']$. Hence, $DegSim(r^{\perp},r'') < DegSim(r^{\perp},r')$ which is absurd since r' has the minimal degree of similarity with $r^{\perp}$ in s^r.*

3.2 SKYRULE Algorithm

Based on the formalization, we proposed the SKYRULE algorithm allowing to discover undominated rules. In SKYRULE algorithm we use the following variables for accumulating data during the execution of the algorithm:

- The variable Sky: is a variable initialized to empty set and it is used to contain the undominated rules.
- The variable C: is a variable containing the set of all current candidate rules to be qualified as undominated; it is initialized to $\mathcal{R}$.
- The variable $\mathcal{E}$: is a variable containing all current set covering the undominated space of all undominated rules; it is initialized to $\mathcal{R}$ because initially, all rules are considered undominated.

Algorithm 1: SKYRULE

Input: $\Omega = (\mathcal{R}, \mathcal{M})$
Output: $Sky_M(\Omega)$: set of undominanted rules of Ω.

```
1  begin
2    Sky ← ∅
3    C ← R
4    E ← {R}
5    while C ≠ ∅ do
6      r* ← r ∈ C having min(DegSim(r,r⊥))
7      C ← C\{r*}
8      for i=1 to k do
9        s_i^{r*} ← ∅
10     Sky ← Sky ∪ {r*}
11     foreach e ∈ E such that r* ∈ s do
12       foreach r ∈ e do
13         if r* ≻ r then
14           C ← C\{r}
15         else
16           for i=1 to k do
17             if r[m_i] ≻ r*[m_i] then
18               s_i^{r*} ← s_i^{r*} ∪ {r}
19       E ← E\{e}
20     E ← E ∪ {s_1^{r*}, ..., s_k^{r*}}
21   return Sky
22 end
```

Informally, the algorithm works as follows:

- If the set of candidate rules C is empty, then the algorithm terminates and all undominated rules are in Sky.
- Otherwise, each rule r in C might be an undominated rule. If r has the minimal degree of similarity with the reference rule $r^{\perp}$ then, r is an undominated rule and it is added to Sky (*i.e.,* r is no longer candidate and it is deleted from C). After that, only the undominated space containing r is explored as follows: for each rule r' in this undominated space r' is compared with r, then we have two cases:
  1. if r' is dominated by r, then r is no longer candidate and it is deleted from C.
  2. otherwise, r' is not dominated by r, *i.e.,* r' is still a candidate rule and it is added to the undominated subspace of r according to definition 6.

  Then, the undominated space containing r is deleted from $\mathcal{E}$ and the undominated space of r is added to $\mathcal{E}$. This process is repeated until there is no more candidate left.

4 RANKING ASSOCIATION RULES

The SKYRULE algorithm allows identifying the undominated rules which are supposed to be the most relevant ones. However, this output might not answer a personalized user query. Indeed, the user often need a specified number of relevant rules which may be more or less than what SKYRULE generates. In the first case *i.e.,* the user asks for a subset of the undominated rules, a selection is required among the SKYRULE output. Since, SKYRULE generate only relevant rules, the most relevant among them must be returned to the user. This selection cannot be performed unless a ranking has be done within the undominated rules. In the second case *i.e.,* the user asks for a set of relevant rules larger than the set of undominated rules, the rules that must be added to the SKYRULE output are necessarily a part from the set of dominated rules. The composition of this part requires a selection among all the dominated rules. This selection cannot be performed unless a ranking has be done within the dominated rules. Hence, a ranking process must be performed on the whole set of rules.

In this section, we present our second contribution: we show that we can perform a comprehensive ranking using SKYRULE. For this purpose, we give the two following objective conditions:

1. Any dominated rule cannot be better ranked than an undominated one.
2. Two undominated rules must be ranked based on degree of similarity with a reference rule.

4.1 Succession relationship

In this section, we introduce the notion of *succession relationship*. This notion is based on the dominance relationship. First, we define it at the level of rules. Then we define it at the level of rule sets. The two definitions are essentiel to state Lemma 4. That lemma puts the corner stone of our approach that uses the skyline operator to establish a ranking process. This process is described by RANKRULE (see algorithm 2).

Definition 7 ***(Successor rule)*** *Let two rules r, $r' \in \mathcal{R}$, we say r succeed r', noted by $r \lhd r'$ iff $r' \succ r$ and $\nexists\, r''$ such that $r' \succ r'' \succ r$.*

Example. Consider the relation table Ω in Table1(b), $r_6 \lhd r_4$ but $r_5 \not\lhd r_4$ since $r_4 \succ r_6 \succ r_5$.

Definition 8 ***(Succession Operator)*** *Let E be a set of rules such that $E \subseteq \mathcal{R}$. The successeur set of E in $\mathcal{R}$ with respect to $\mathcal{M}$ is defined as follows: $Succ_{\mathcal{M}}(E,\mathcal{R}) = \{\, r \in \mathcal{R} \setminus E \mid \exists\, r' \in E,\, r \lhd r' \wedge \nexists\, r'' \in E,\, (r'' \succ r \wedge r \not\lhd r'')\}$*

Example. Let's consider our running example using the relation table Ω in Table1(b) and suppose $E = \{r_1, r_4\}$, $r_1 \succ r_3 \succ r_2$, $r_1 \succ r_5 \succ r_7$, $r_5 \succ r_8$ and $r_4 \succ r_6 \succ r_5$ then $Succ_{\mathcal{M}}(E,\mathcal{R}) = \{r_3, r_6\}$. Notice that, although $r_5 \lhd r_1$, $r_5 \notin Succ_{\mathcal{M}}(E,\mathcal{R})$ because $r_5 \not\lhd r_4$.

Lemma 4 *Given a set of rules E, one has the following relation:*

$$Succ_{\mathcal{M}}(Sky_M(E),E) = Sky_M(E \backslash Sky_M(E))$$

Proof 4 *Let E be a set of rules:*

1. First we show that $Succ_{\mathcal{M}}(Sky_M(E),E) \subseteq Sky_M(E \backslash Sky_M(E))$:
Given $r \in Succ_{\mathcal{M}}(Sky_M(E),E)$ then $r \in E \backslash Sky_M(E)$. For all $r' \in Sky_M(E)$, we have two cases :

- *If $r' \succ r$, then $r \lhd r'$ which means $\nexists\, r'' \in E \backslash Sky_M(E)$ such that $r' \succ r'' \succ r$.*
- *If $r' \not\succ r$, then $\nexists\, r''$ in $E \backslash Sky_M(E)$ such that $r' \succ r''$ and $r'' \succ r$*

Thus r cannot be dominated by any rule in $E \backslash Sky_M(E)$ i.e., $r \in Sky_M(E \backslash Sky_M(E))$.

2. Second we show that $Succ_{\mathcal{M}}(Sky_M(E),E) \supseteq Sky_M(E \backslash Sky_M(E\ ))$:
*Given $r \in Sky_M(E \backslash Sky_M(E\ ))$ then $\nexists\, r' \in E \backslash Sky_M(E)$ such that $r' \succ r$ **(a)**. Moreover, as $r \in E \backslash Sky_M(E)$ then $\exists\, r'' \in Sky_M(E)$ such that $r'' \succ r$ **(b)**. Thus **(a)** and **(b)** mean that $r \lhd r''$ **(c)**.*
*Furthermore, we suppose that $\exists\, r' \in Sky_M(E)$ such that $r_1 \succ r$ and $r \not\lhd r_1$, then $\exists\, r_2 \in E \backslash Sky_M(E)$ such that $r_1 \succ r_2 \succ r$ which is absurd (see **(a)**). Thus, $\nexists\, r_2 \in E \backslash Sky_M(E)$ such that $r_1 \succ r_2 \succ r$ **(d)**. Hence, according to **(c)** and **(d)**, r belongs to $Succ_{\mathcal{M}}(Sky_M(E),E)$.*

Algorithm 2: RANKRULE

Input: $\Omega = (\mathcal{R}, \mathcal{M})$
Output: Ordered sets of ordered rules

```
1 begin
2     p ← 0
3     while R ≠ ∅ do
4         p ← p + 1
5         E_p ← SKYRULE(Ω)
6         R ← R \ E_p
7         Ω ← (R, M)
8     return (E_1, . . ., E_p)
9 end
```

Example. In this example, we apply RANKRULE on Ω of Table 1. Since r_1 and r_4 are undominant rules then $E_1 = \{r_1, r_4\}$. Now we ignore r_1 and r_4, the rules which are not dominated are r_3 and r_6. In fact, r_3 is dominated by only r_1 and r_6 is dominated by only r_1, then $E_2 = \{r_3, r_6\}$. Now we ignore also r_3 and r_6, the rules which are not dominated are r_2 and r_5. In fact, r_2 is dominated by r_3 and r_5 is dominated by only r_6, then $E_3 = \{r_2, r_5\}$. Finally, $E_4 = \{r_7, r_8\}$. This example is illustrated by Figure 4.1. The arrow indicates the process direction starting from the undominated rules. E_1 contains the top ranked rules which are them selves ranked within E_1 from left to right based on $DegSim$: r_1 is better ranked than r_4.

4.2 Duality

RANKRULE performs ranking by starting from the set of the most relevant rules (*i.e.,* the undominated rules) and uses it to identify the next ranked set (*i.e.,* the successor). Meanwhile, another dual perspective remains possible. It relies on starting from the set of the less relevant rules (*i.e.,* rules that do not dominate other rules) and using them to identify the previous ranked rule set that we called *predecessor* set. We do not give a formalization of this dual perspective, yet we explain how it works by the following illustrative example.

Figure 1. The output of RANKRULE applied on Ω.

Example. We consider Ω of Table 1. First we identify the set of rules which do not dominate any other rules. These rules are r_2, r_7 and r_8 then $E_4 = \{r_2, r_7, r_8\}$. Now ignore these rules. The rules which do not dominate any other rules are r_3 and r_5. In fact, r_3 dominates only r_2 and r_5 dominates only r_7 and r_8, then $E_1 = \{r_3, r_5\}$. Now we ignore also r_3 and r_5, The rules which do not dominate any other rules are r_1 and r_6 since they dominate r_3 and r_5 respectively, then $E_2 = \{r_1, r_6\}$. Finally, $E_1 = \{r_4\}$.

Figure 2. The dual RANKRULE applied on Ω.

5 CONCLUSION

In this paper we proposed an approach that addresses the problem of rule selection and ranking. This approach is not hindered by the abundance of measures which is the issue of several works. These works have been devoted to measure selection in order to find one best measure, whereas the real issue lies in selecting and ranking rules to help with decision making. We proposed two algorithms SKYRULE and RANKRULE to perform these two tasks based on the dominance relationship. When using our algorithms, the user does not have to

worry neither about the heterogeneity of measures nor about specifying thresholds. As future works, we plan to formalize the dual of RANKRULE and to find the relationship between them that allows to obtain the output of one of them from the output of the other. Another importante prospective is to study the impact of any change within the relational table Ω, such as the insertion of new measures of new rules, on the ranking or the selection.

REFERENCES

[1] R. Agrawal and R. Skirant, 'Fast algorithms for mining association rules', in *Proceedings of the 20th Intl. Conference on Very Large Databases, Santiago, Chile*, pp. 478–499, (June 1994).

[2] J. Blanchard, F.Guillet, H.Briand, and R.Gras, 'Assessing the interestingness of rules with a probabilistic measure of deviation from equilibrium.', in *The XIth International Symposium on Applied Stochastic Models and Data Analysis, Brest, France*, pp. 191–200, (2005).

[3] S. Borzsony, D. Kossmann, and K. Stocker, 'The skyline operator.', in *ICDE*, pp. 421–430, (2001).

[4] L. Geng and H.J Hamilton, 'Choosing the right lens : Finding what is interesting in data mining', in *Quality Measures in Data Mining, ISBN 978-3-540-44911-9*, pp. 3–24, (2007).

[5] M.J. Heravi and O. Zaiane, 'A study on interestingness measures for associative classifiers., acm sac'10', pp. 1039–1046, (2010).

[6] R. J. Hilderman and H. J. Hamilton, 'Knowledge discovery and measures of interest', in *volume 638 of The International Series in Engineering and Computer Science*, (2001).

[7] R. J. Hilderman and H. J. Hamilton, 'Measuring the interestingness of discovered knowledge: A principled approach', in *Intelligent Data Analysis 7*, pp. 347–382, (2003).

[8] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo, 'Finding interesting rules from large sets of discovered association rules', in *Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM'94),ACM Press*, pp. 401–407, (November 1994).

[9] H.T. Kung, F. Luccio, and F.P. Preparata, 'On finding the maxima of a set of vectors', in *J. ACM, vol. 22, no. 4*, pp. 469–476, (1975).

[10] F. Lenca, P. Meyer, P. Picouet, B. Vaillant, and S. Lallich, 'Critères d'évaluation des mesures de qualit en ECD', in *Revue des Nouvelles Technologies de l'Information (Entreposage et Fouille de donnes), (1)*, pp. 123–134, (2003).

[11] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich, 'On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid'.

[12] M. Maddouri and J. Gammoudi, 'On semantic properties of interestigness measures for extracting rules from data', in *ICANNGA (1), Springer-Verlag*, pp. 148–158, (2007).

[13] H. Manilla, 'Methods and problems in data mining', in *Proceedings of the 6th biennial Intl. Conference Database theory (ICDT'97), LNCS,V ol. 1186, Springer-Verlag*, pp. 41–55, (January 1997).

[14] J. Matousek, 'Computing dominances in En.', in *Inform. Process. Lett., 38*, pp. 227–278, (1991).

[15] M. Ohsaki, Y. Sato, S. Kitaguchi, and H. Yokoi, 'Comparison between objective interestingness measures and real human interest in medical data mining.', in *Orchard, R., Yang, C., Ali, M., eds.: 17th international conference on Innovations in Applied Artificial Intelligence (IEA/AIE 2004). Volume 3029 of Lecture Notes in Artificial Intelligence., Springer-Verlag*, pp. 1072–1081, (2004).

[16] R.E. Steuer, '*Multiple Criteria Optimization: Theory, Computation and Application.*', in *John Wiley, 546*, (1986).

[17] P. Tan and V. Kumar, 'Interestigness measures for association patterns: A perspective', in *Proceedings of Workshop on Postprocessing in Machine Learning and Data Mining*, (2000).

[18] P. Tan, V. Kumar, and J. Srivastava, 'Selecting the right interestingness measure for association patterns', in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ICDM'02), ACM Press*, pp. 32–41, (2002).

[19] B. Vaillant, F. Lenca, and S. Lallich., 'A clustering of interestingness measures.', in *Discovery Science. Volume 3245 of Lecture Notes in Artificial Intelligence., Springer-Verlag*, pp. 290–297, (2004).